OXFORD
UNIVERSITY PRESS

Archives of Clinical Neuropsychology 31 (2016) 313–331

Archives of CLINICAL NEUROPSYCHOLOGY

False-Positive Error Rates for Reliable Digit Span and Auditory Verbal Learning Test Performance Validity Measures in Amnestic Mild Cognitive Impairment and Early Alzheimer Disease

David W. Loring[1,2,*], Felicia C. Goldstein[1], Chuqing Chen[3], Daniel L. Drane[1,2], James J. Lah[1], Liping Zhao[4], Glenn J. Larrabee[5], for the Alzheimer's Disease Neuroimaging Initiative[†]

[1]*Department of Neurology, Emory University School of Medicine, Atlanta, GA, USA*
[2]*Department of Pediatrics, Emory University School of Medicine, Atlanta, GA, USA*
[3]*Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, USA*
[4]*Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, GA, USA*
[5]*Independent Practice, Sarasota, FL, USA*

*Corresponding author at: Emory University Brain Health Center, 12 Executive Park, Atlanta, GA 30329, USA.
E-mail address: dloring@emory.edu (D.W. Loring).

Accepted 18 February 2016

Abstract

Objective: The objective is to examine failure on three embedded performance validity tests [Reliable Digit Span (RDS), Auditory Verbal Learning Test (AVLT) logistic regression, and AVLT recognition memory] in early Alzheimer disease (AD; $n = 178$), amnestic mild cognitive impairment (MCI; $n = 365$), and cognitively intact age-matched controls ($n = 206$).
Method: Neuropsychological tests scores were obtained from subjects participating in the Alzheimer's Disease Neuroimaging Initiative (ADNI).
Results: RDS failure using a $\leq$7 RDS threshold was 60/178 (34%) for early AD, 52/365 (14%) for MCI, and 17/206 (8%) for controls. A $\leq$6 RDS criterion reduced this rate to 24/178 (13%) for early AD, 15/365 (4%) for MCI, and 7/206 (3%) for controls. AVLT logistic regression probability of $\geq$.76 yielded unacceptably high false-positive rates in both clinical groups [early AD = 149/178 (79%); MCI = 159/365 (44%)] but not cognitively intact controls (13/206, 6%). AVLT recognition criterion of $\leq$9/15 classified 125/178 (70%) of early AD, 155/365 (42%) of MCI, and 18/206 (9%) of control scores as invalid, which decreased to 66/178 (37%) for early AD, 46/365 (13%) for MCI, and 10/206 (5%) for controls when applying a $\leq$5/15 criterion. Despite high false-positive rates across individual measures and thresholds, *combining* RDS $\leq$ 6 and AVLT recognition $\leq$9/15 classified only 9/178 (5%) of early AD and 4/365 (1%) of MCI patients as invalid performers.
Conclusions: Embedded validity cutoffs derived from mixed clinical groups produce unacceptably high false-positive rates in MCI and early AD. Combining embedded PVT indicators lowers the false-positive rate.

Keywords: Performance validity test; False-positive rate; Test specificity; Reliable Digit Span; AVLT

Introduction

The inclusion of formal performance validity tests (PVTs) as routine components of neuropsychological assessment protocols is a recommended practice of contemporary neuropsychology (Bush et al., 2005). PVTs may be either stand-alone measures that are included explicitly to evaluate validity alone, or embedded measures in which validity scores are derived from existing neuropsychological measures of motor function, attention, memory, or problem solving that represent clinically atypical performance

[†] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

doi:10.1093/arclin/acw014 Advance Access publication on 15 April 2016

(e.g., poorer recognition memory relative to free recall). Embedded PVTs do not require additional testing and are therefore more easily applied in non-forensic evaluations in which time and personnel resources may be limited.

Compared with other areas in clinical neuropsychology, PVT research is encumbered by significant methodological challenges since no external independent standard exists against which criterion-related validity can be established. Two research designs address this problem (Rogers, 1997), although these approaches are also associated with important limitations. The first design involves simulation, and contrasts performance of non-injured subjects who are asked to simulate acquired cognitive impairment from brain injury to that of patients with independently established clinical disease who are not in litigation or other compensation-seeking actions. The second approach, the known-group design, relies on external criteria of malingering for subject classification and contrasts performance of litigants meeting malingering criteria with that of non-litigating clinical cases without evidence to indicate malingering. The most frequently used criteria for defining malingering for known-group designs were presented by Slick, Sherman, and Iverson (1999). Because of the risk of mischaracterization in different clinical populations, Slick and colleagues (1999) criteria require that PVT failure cannot be the primary result of neurologic, psychiatric, or developmental impairments when inferring malingered neurocognitive impairment (Criterion D) to minimize false-positive identification.

False-positive rate is critical for establishing the accuracy of all diagnostic tests including those to infer malingering (Larrabee, 2012; Straus, Richardson, Glasziou, & Haynes, 2010). Positive predictive power (i.e., the probability of having the condition of interest) is defined by both true positive and false-positive test classifications, and consequently is greater when false-positive rates are low. Based on reviews of PVT research in medicolegal contexts, PVT investigators strive to maintain false-positive PVT rates at 10% or less (Boone, 2013). To enhance generalizability across groups while maintaining low false-positive error rates, clinical comparator groups often include non-litigating patients who have suffered moderate or severe traumatic brain injury (TBI) (Larrabee, 2003; Wolfe et al., 2010). By including groups with unequivocal impairment and establishing cutoffs to minimize false-positive rates in these groups, the goal is to maintain low false-positive rates on subsequent clinical PVT application. Approaches to minimize false-positive PVT error rates include adjusting cut-off scores based upon clinical diagnosis and determining whether sufficient cognitive ability exists on various neuropsychological measures to adequately perform a specific PVT task (Larrabee, 2014).

Reliable Digit Span (RDS) is an embedded PVT that reflects performance consistency across both trials of each digit span length (Greiffenstein, Baker, & Gola, 1994) and has been investigated in several dementia series. Using an RDS criterion of $\leq$7 to infer performance invalidity in 20 patients with probable AD (NINCDS-ADRA criteria; average MMSE = 22.2/30), only 6/20 patients (30%) were classified as having valid scores (Merten, Bossink, & Schmand, 2007). An additional RDS concern is that only 9/14 (64%) cognitively intact control subjects obtained valid scores, thus reflecting a potential age confound when using this criterion. A lower RDS criterion of $\leq$6 in a much larger (n = 1336) but more heterogeneous clinical sample (e.g., TBI, stroke, multiple sclerosis, Parkinson disease, lupus, cerebral palsy, learning disability, academic problems) resulted in an RDS specificity of 1029/1336 (77%), which improved to 1189/1336 (89%) when applying a $\leq$5 RDS criterion (Heinly, Greve, Bianchini, Love, & Brennan, 2005). Clinical diagnoses with the highest false-positive classification using the $\leq$6 RDS criterion included stroke (155/517; 30%) and memory impaired (73/228; 32%), with values dropping to 72/517 (14%) and 41/228 (18%), respectively, when applying a $\leq$5 RDS threshold.

In a mixed dementia cohort, an RDS criterion of $\leq$6 yielded specificities of 38/44 (86%) in patients with the mean MMSE = 23.5/30, only 18/30 (60%) in subjects with the mean MMSE = 17.6/30, with a further decline to only 2/9 (22%) in patients with the mean MMSE = 9.4/30 (Dean, Victor, Boone, Philpott, & Hess, 2009). Although these data establish elevated false-positive risk that increases with dementia severity, unfortunately, MMSE scores were available only for slightly more than half the total sample size of 214 subjects. When analyzed according to dementia diagnosis, sample sizes were small—RDS specificity was 23/31 (74%) in early AD, 15/26 (58%) in vascular dementia, and 27/36 (75%) in frontotemporal dementia.

These studies not only raise serious concerns for generalizing RDS classification criteria to patients with dementia, but also have important limitations including small sample size and heterogeneity of clinically referred samples. These concerns have been partially addressed in a retrospective series of 142 patients with probable AD diagnosed by NINCDS-ADRDA criteria referred from a university-based memory disorders program (Kiewel, Wisdom, Bradshaw, Pastorek, & Strutt, 2012). A wide range of dementia severity was studied, with MMSE scores as low as 1/30 included. For mild AD (mean MMSE = 23.4/30), a $\leq$6 RDS criterion was associated with a false-positive error rate of 9/78 (12%). The false positives for moderate AD (mean MMSE = 16.8/30) were 10/41 (24%), and increased to 19/23 (83%) for severe AD (mean MMSE = 7.7/30). Thus, even an RDS $\leq$6 threshold resulted in unacceptably high levels of false-positive classification, particularly in patients with more severe dementia.

Other embedded/derived PVTs have been developed for common neuropsychological tests including the Rey Auditory Verbal Learning Test (AVLT). Because recognition memory is often relatively preserved in many neurological illnesses, AVLT recognition has been demonstrated to have potential PVT utility in several reports. Compensation seeking mild TBI patients failing a forced choice PVT not only performed more poorly on AVLT recognition (n = 24; 8.4/15) than a similar patient group passing PVT (n = 17; 12.4/15), but also obtained lower scores than brain-injured subjects not seeking compensation (n = 68; 11.6/15)

(Binder, Villanueva, Howieson, & Moore, 1993). An AVLT recognition score of ≤5 correctly classified 20/75 (27%) of all mild TBI patients without respect to external PVT performance, but more importantly, misclassified only 4/80 (5%) of the brain injury group. However, AD patients were excluded from the brain injury group composition.

In their review of AVLT recognition studies in which incentives for poor performance were present, Boone, Lu, and Wen (2005) reported average AVLT recognition ranging from 6.8/15 to 9.9/15. In their own dataset, the PVT fail/compensation-seeking group ($n = 61$) averaged 7.7/15 versus 12.9/15 for clinical patients ($n = 88$), and averaged 13.0/15 for healthy controls ($n = 25$). An AVLT recognition cut-score of ≤9/15 yielded a sensitivity of 67% and specificity of 93% for characterizing patients with suspect effort compared with controls and clinical patients combined. Again, patients with AD were not included in the clinical comparison group, and in addition, patients with generalized cognitive impairment as reflected by Full Scale IQ < 70 were also excluded.

In a patient series with mixed neurologic diagnoses seeking compensation and classified according to multiple stand-alone PVTs as credible ($n = 112$) or non-credible ($n = 63$), an AVLT recognition score of ≤9/15 was associated with a sensitivity of 48% and specificity of 91% (Whitney & Davis, 2015). As with previous reports, patients diagnosed with dementia were excluded from these analyses.

In a sample of compensation-seeking TBI patients who either failed ≥2 PVTs ($n = 62$) or passed all PVTs ($n = 68$), two AVLT variables utilizing logistic regression with Bayesian Model Averaging were identified that accurately classified patients (Davis, Millis, & Axelrod, 2012): total words recalled over the five learning trials, and the AVLT recognition score. Logistic regression yielded an area under the receiver operating curve of 0.85 demonstrating excellent discrimination; a cutting score of ≥0.70 yielded a sensitivity of 55% with a specificity of 91%. The PVT pass group averaged 12.5/15 on AVLT recognition versus an average of 9.2/15 for the PVT fail group. Patients with dementia or mental retardation were excluded.

AD is associated with defective encoding that is reflected in impaired recognition memory. For example, although Parkinson disease dementia patients ($n = 12$) had comparable performance on AVLT recognition compared with 38 control subjects (13.1/15 vs. 13.6/15), AVLT recognition scores were significantly lower for 18 moderate AD (10.8/15) and for 33 severe AD patients (8.1/15) (Tierney et al., 1994). Consequently, the omission of AD patients from clinical group composition when deriving classification statistics will decrease the likelihood of false-positive AVLT recognition errors resulting in higher reported specificity.

There are important limitations of these reports in addition to small sample sizes. First, the subjects used in PVT research are typically derived from clinically referred samples of convenience in which referral biases/spectrum biases may influence the sample representativeness. Second, patients with moderate to severe dementia in which the diagnosis of dementia is not in doubt makes Criterion D of Slick and colleagues relevant. Third, with the exception of the small series reported by Merten and colleagues (2007), these reports described patients who were retrospectively identified and did not include a cognitively intact control comparison group. Fourth, because patients with dementia or low Full Scale IQ were excluded from embedded AVLT measures, classification patterns in these conditions remain unknown and generalization to these populations may be inappropriate. Fifth, there have been no studies to date examining false-positive rates in amnestic mild cognitive impairment (MCI), which is considered to be prodromal AD and represents an intermediate stage between normal cognitive function and full dementia expression. The performance of MCI patients is critical to characterize since MCI represents a patient population for whom clinical neuropsychological testing often provides the greatest diagnostic clarity (Bondi & Smith, 2014).

The present report evaluates PVT false-positive rates of three embedded PVTs in early AD, amnestic MCI (single domain or multi-domain), and cognitively intact controls. The three PVTs include: (i) RDS, (ii) Rey AVLT logistic regression (Davis et al., 2012), and (iii) Rey AVLT recognition memory (Binder et al., 1993). Subjects include a large sample of clinical research volunteers ($n = 749$) enrolled in the Alzheimer's Disease Neuroimaging Initiative (ADNI). We hypothesized that the frequency of embedded PVT failure for each PVT measure would vary across diagnostic groups, reflecting the effects of differences in disease severity. We also explore rates of PVT classification using different thresholds at different levels to establish potential cutpoints that are not associated with high levels of false-positive classification, and evaluate change in failure rate for PVT combinations of PVTs (i.e., RDS and AVLT recognition). We describe the relationship of various demographic and neuropsychological factors influencing PVT performances in each of the three subject groups. Finally, we present classification accuracy based upon cognitive factors associated with PVT failure including level of performance on the MMSE, Trail Making Part B, and AVLT delayed free recall.

Materials and Methods

Study Participants

There were 178 subjects diagnosed with early AD, 365 subjects with amnestic MCI, and 206 cognitively intact controls. Subjects were enrolled in the ADNI, a multi-center, 3-year longitudinal investigation to identify structural and functional brain changes including biomarkers predictive of MCI conversion and AD progression. ADNI consists of 59 clinical recruiting sites

across the United States, and subjects were recruited from specialty memory clinics, from Alzheimer Disease Research Center (ADRC) registries, and through advertisements placed in local media. All subjects provided written informed consent.

Subjects were in the first data collection series conducted from 2005 to 2009 (ADNI1), which contained item level performance on most cognitive measures. Inclusion criteria for ADNI1 entailed an age range between 55 and 90 years old, a minimum of 6 years of formal education, fluency in English or Spanish, Hachinski Ischemic Scale (Hachinski et al., 1975) scores $\leq$4/18, and Geriatric Depression Scale Short Form scores $<$6/15 points. Subjects were excluded if they were taking medications with anticholinergic properties (e.g., diphenhydramine, amantadine), regular narcotic analgesic (e.g., oxycodone), antiparkinsonian medications (e.g., levodopa), or sedatives/benzodiazepines (e.g., clonazepam).

Participants were classified as cognitively intact controls, amnestic MCI, or early AD based upon research criteria that included the Mini-Mental State Examination (MMSE) (Folstein, Folstein, & McHugh, 1975), immediate and delayed recall of the first Logical Memory story (Anna Thompson) from the Wechsler Memory Scale-Revised (Wechsler, 1987), and the Clinical Dementia Rating (CDR) interview conducted with each participant's partner (Morris, 1993). Of 814 subjects enrolled in ADNI1, 749 were administered Digit Span and AVLT and were included in this study.

Cognitively intact controls were defined as having no significant memory complaints beyond those expected for age, a normal education adjusted cutoff score on Logical Memory delayed recall (Aisen et al., 2 010), an MMSE score between 24 and 30/30 points, a CDR score of 0/3 (including a 0 on the Memory Box score; Hughes, Berg, Danziger, Coben, & Martin, 1982), and intact instrumental activities of daily living. MCI subjects had a memory complaint or a memory problem that was noted by their partner, an abnormal education adjusted cutoff score on Logical Memory, a MMSE score between 24 and 30/30, a CDR score including the Memory Box score of 0.5/3, and relatively preserved instrumental activities of daily living. MCI patients were either amnestic MCI, or multi-domain MCI that included memory as one of the affected domains. Finally, participants with early AD had a memory complaint or memory problem that was noted by the study partner, an abnormal education adjusted cutoff score on Logical Memory delayed recall, an MMSE score between 20 and 26/30, a CDR score between 0.5 and 1.0/3, and met NINCDS/ADRDA criteria for probable AD. Three subjects in the present study had MMSE scores $>$26/30 and were reclassified as having early AD after the screening visit based upon consensus conference. A complete list of additional inclusion/exclusion criteria for ADNI1 is available at http://www.adni-info.org/Scientists/doc/ADNI_GeneralProceduresManual.pdf.

Neuropsychological Tests

American National Adult Reading Test. Premorbid Verbal IQ was estimated using the American National Adult Reading Test (AmNART), which consists of reading 50 words with atypical grapheme to phoneme relationship (Grober & Sliwinski, 1991). The number of pronunciation errors is used for regression-based VIQ estimation.

Boston Naming Test. Visual confrontation naming was assessed using a 30-item version of the Boston Naming Test (BNT) (Kaplan, Goodglass, & Weintraub, 1983). To estimate normative performance, the obtained score (spontaneous and semantic cue) was doubled, and age-scaled scores derived using Mayo's Older American Normative Studies (MOANS) (Steinberg, Bieliauskas, Smith, Langellotti, & Ivnik, 2005).

Trail Making Parts A and B. Visual scanning speed was determined by performance on Trail Making Parts A and B (Army Individual Test Battery, 1944). The maximum time allowed for Trail Making Part A was 150 s and the maximum time allowed for Trail Making Part B was 300 s. Age-scaled scores were derived from MOANS (Steinberg, Bieliauskas, Smith, & Ivnik, 2005).

Digit Symbol Substitution Test. This measure of graphomotor processing speed is from the WAIS-R and is the forerunner to the Coding subtest used in current versions of Wechsler Intelligence Scales (Wechsler, 1981). In contrast to Coding, however, Digit Symbol Substitution is timed at 90 s. Age-scaled scores were derived from MOANS (Ivnik et al., 1992a, 1992b).

Animal naming/category fluency. Generative verbal fluency to the prompt to list as many animals as possible in 60 s was obtained without further elaboration of instructions. Normative performance was obtained using regression estimation (Mitrushina, Boone, Razani, & D'Elia, 2005, p. 764).

Digit span. Forward and backward digit span were assessed using the version presented in the WAIS-R (Wechsler, 1987). Age-scaled standard scores were derived from MOANS (Ivnik et al., 1992a, 1992b).

Rey Auditory Verbal Learning Test. This is a serial word list learning task presenting 15 words over 5 trials (Rey, 1941). A distractor list is presented for a single trial, followed by spontaneous recall of the initial 15 words. Following a 30 min delay, free recall of the

original word list is obtained followed by recognition (Ivnik et al., 1992a, 1992b) using the "Form AB" 30 word recognition list (Schmidt, 1996, p. 76). The recognition procedure consists of 15 target words combined with 15 foils presented on a single sheet, with subjects instructed to circle all words remembered from the original list. Normative values for learning over trials (LOT), delayed free recall, and delayed recognition memory were obtained using norms from the MOANS (Ivnik et al., 1992a, 1992b).

PVT Measures

Reliable Digit Span. Digit Span was administered using standard instructions, with both trials of each span length scored (Greiffenstein et al., 1994). The maximum reliable span lengths for the forward and the backward repetition conditions (i.e., scores of 1 on both trials of the same span length prior to discontinuation of the subtest) were summed to form a composite RDS value.

Logistic regression. Logistic regression estimates were derived from AVLT 5 trial learning sum (not Mayo's Learning Overt Trials, LOT) and recognition performance using the following formula: Probability of performance invalidity $= (e^{[6.61-(\text{AVLT Total})-\text{Recognition}\times 0.258)]})/1 + e^{[6.61-(\text{AVLT Total})-(\text{Recognition}\times 0.258)]}$ (Davis et al., 2012).

AVLT recognition. Delayed AVLT recognition memory was obtained by presenting 15 target items with 15 distractor items (Ivnik et al., 1992a, 1992b) and asking the subject to circle items remembered from the original 5 trial learning list (Binder et al., 1993). The number of correct recognitions without any correction for number of false-positive intrusions was used for classification.

Results

Subjects

There were 434 (58%) males and 315 (42%) females; 697 (93%) were white, and 460 (61%) participants had a college degree or higher. The average age was 75.7 years ($SD = 7.5$) for the early AD group, 74.9 years ($SD = 7.2$) for MCI, and 76.0 years ($SD = 5.0$) for cognitively intact controls. The average MMSE was 23.3/30 ($SD = 2.0$) for early AD, 27.0/30 ($SD = 1.8$) for MCI, and 29.1/30 ($SD = 1.0$) for controls. Other demographic variables and levels of neuropsychological performance are included in Table 1. Significant group differences across all neuropsychological variables were present, with effect sizes ranging from 0.06 (AmNART Verbal IQ) to 0.73 (Anna Thompson delayed recall), and all pairwise contrasts using the Bonferroni correction were significant across all reported neuropsychological measures.

Reliable Digit Span

Group results. The average RDS score was 8.4 ($SD = 1.9$) for early AD subjects, 9.5 ($SD = 2.0$) for MCI subjects, and 10.3 ($SD = 2.0$) for controls. The values significantly differed across groups using one-way ANOVA demonstrating a clear disease relationship on RDS scores ($p < .0001$; partial $\eta^2 = 0.10$).

Individual classification. Individual characterization of RDS performance employed RDS cutpoints for ≤ 7, ≤ 6, and ≤ 5 applied to early AD, MCI, and control groups (see Table 2). The RDS ≤ 7 criterion classified 60/178 (34%) AD patients and 52/365 (14%) MCI patients but only 17/206 (8%) controls as performing in the invalid range. Lowering the threshold to RDS ≤ 6 decreased false-positive classification in all three groups, and although the false-positive rate in both MCI and controls fell below 5%, false-positive rate remained elevated for AD (24/178; 13%). It was not until a ≤ 5 RDS criterion was applied that the false-positive rate fell below 10% (6/178; 3%).

Significant differences in the frequency of invalid characterization were seen across all three cutpoints (see Table 2). The greatest difference in classification rates across the three groups was present with the RDS ≤ 7 cutpoint. However, even using the RDS ≤ 5 criterion that minimizes false-positive errors in early AD, there was a significant group difference in false-positive frequency.

When performing pairwise group contrasts, comparison of controls and MCI using the RDS ≤ 7 cutpoint demonstrated significant differences in invalid characterization ($\chi^2 = 4.5$, $p = .04$, $\eta = 0.09$). There were also invalid characterization differences between MCI and AD subjects ($\chi^2 = 27.7$, $p < .0001$, $\eta = 0.23$).

We performed similar pairwise group follow-up comparisons using the RDS ≤ 6 cutpoint. In contrast to the analysis using the RDS ≤ 7 criterion, there was no significant difference in the pass/fail frequencies of total RDS scores between the control and MCI groups ($\chi^2 = 0.2$, NS). However, the difference between early AD and MCI groups remained significant ($\chi^2 = 15.8$, $p < .0001$, $\eta = 0.17$) indicating differential false-positive classification rate associated with the specific diagnosis.

Table 1. Neuropsychological performance across groups

Test	Early AD ($n = 178$)	MCI ($n = 365$)	Controls ($n = 206$)	Partial η^2
AmNART VIQ	113.7 (10.0)	115.8 (9.8)	120.1 (8.3)	0.06
Digit Span SS	9.7 (3.0)	11.2 (3.0)	12.6 (3.0)	0.10
AVLT LOT SS	6.2 (2.6)	8.0 (3.1)	11.3 (3.0)	0.29
AVLT 30 min delay (raw)	0.7 (1.7)	2.9 (3.3)	7.4 (3.7)	0.39
AVLT 30 min delay (SS)	5.0 (1.8)	6.8 (3.1)	11.0 (3.4)	0.38
AVLT Recognition (raw)	7.1 (4.0)	9.8 (3.6)	12.8 (2.8)	0.26
AVLT Recognition (SS)	5.3 (3.0)	7.4 (3.3)	10.6 (2.8)	0.27
Anna Thompson Immediate	4.1 (2.8)	7.1 (3.2)	13.7 (3.5)	0.56
Anna Thompson Delay	1.3 (1.9)	3.9 (2.7)	13.0 (3.6)	0.73
Boston Naming (raw est)	44.3 (12.6)	51.1 (8.1)	56.7 (4.7)	0.18
Boston Naming (SS est)	8.0 (4.2)	10.3 (3.8)	12.8 (3.4)	0.17
Animal Naming	12.4 (5.0)	16.0 (4.9)	20 (5.7)	0.22
Animal Naming (SS)	7.0 (3.2)	9.2 (3.1)	11.8 (3.4)	0.22
Trail Making Part A (s)	66.9 (36.7)	44.0 (22.0)	36.6 (13.4)	0.17
Trail Making Part A (SS)	7.2 (3.5)	10.0 (3.2)	11.3 (2.7)	0.18
Trail Making Part B (s)	188.4 (95.6)	128.6 (72.5)	89.0 (44.0)	0.19
Trail Making Part B (SS)	6.0 (3.9)	9.3 (3.7)	11.7 (2.8)	0.24
Digit Symbol (raw, 90 s)	26.3 (13.3)	36.9 (11.3)	45.7 (10.1)	0.27
Digit Symbol (SS)	6.9 (3.6)	9.3 (3.4)	12.1 (2.7)	0.25

Note: ANOVAs across groups for each of the variables are statistically significant at $p < .0001$. All pairwise contrasts within each variable are statistically significant at $p < .0001$ (Bonferroni correction) with the exception of the control vs. MCI contrast for Trail Making Part A (raw score), which was statistically significant at the $p = .002$ level.
AmNART = American National Adult Reading Test; VIQ = Verbal Intelligence Quotients; SS = scaled score; AVLT = Auditory Verbal Learning Test; LOT = Learning over Trials.

Table 2. Classification of performance based on Reliable Digit Span with cutpoints of ≤7, ≤6, ≤5

RDS cutpoint	Early Alzheimer disease ($n = 178$)		Mild cognitive impairment ($n = 365$)		Controls ($n = 206$)		Significance	Effect size
	Valid	Invalid	Valid	Invalid	Valid	Invalid	χ^2	η
≤7	118 (66%)	60 (34%)	313 (86%)	52 (14%)	189 (92%)	17 (8%)	$\chi^2 = 48, p < .0001$	0.24
≤6	154 (87%)	24 (13%)	350 (96%)	15 (4%)	199 (97%)	7 (3%)	$\chi^2 = 22, p < .0001$	0.15
≤5	172 (97%)	6 (3%)	362 (99%)	3 (1%)	206 (100%)	0 (0%)	$\chi^2 = 10, p < .007$	0.11

Because of the small cell frequency using the RDS ≤ 5 criterion, Fisher's exact test was used for pairwise group comparison. With this threshold, there were no differences in classification between the control and MCI groups (p = NS). There was, however, a trend in classification differences between the early AD and MCI groups ($p = .07$).

AVLT Logistic Regression

Group results. AVLT logistic regression classification was performed based upon total learning (sum across trials) and the 30 min delayed recognition (Davis et al., 2012). The average logistic regression probability score for early AD was 0.84 ($SD = 0.18$), for MCI was 0.64 ($SD = 0.27$), for controls was 0.27 ($SD = 0.24$). The values significantly differed across groups using one-way ANOVA ($p < .0001$, partial $\eta^2 = 0.43$).

Individual classification. As can be seen in Table 3, using the probability of test invalidity ≥0.51 (i.e., more likely than not to be invalid) for individual subject classification, there was a significant classification difference across the three groups. Of particular note, however, was the extremely high false-positive classification rate in both clinical groups in which 264/365 (72%) MCI subjects and 166/178 (93%) early AD subjects were identified as having invalid memory scores. Using a more conservative cutoff probability ≥.76 was also associated with a high false-positive error rate in both clinical groups, with 159/365 (44%) MCI subjects and 149/178 (79%) early AD identified as invalid.

We performed pairwise group follow-up comparisons using the 0.51 and 0.76 cutpoints. At the .51 probability level, significant group differences were observed contrasting controls and MCI ($\chi^2 = 175.9, p < .0001, \eta = 0.56$) and contrasting MCI and early AD ($\chi^2 = 31.8, p < .0001, \eta = 0.24$). At the .76 probability criterion, significant group differences remained for both the controls versus MCI comparison ($\chi^2 = 88.8, p < .0001, \eta = 0.39$) and MCI versus early AD contrasts ($\chi^2 = 59.4, p < .0001, \eta = 0.33$).

Table 3. Classification of performance based upon Auditory Verbal Learning Test logical regression model generated probabilities ranging from .51 to .96

	Early Alzheimer disease ($n = 178$)		Mild cognitive impairment ($n = 365$)		Controls ($n = 206$)		Significance	Effect size
Invalidity	Valid	Invalid	Valid	Invalid	Valid	Invalid	χ^2	η
$p \geq .51$	12 (7%)	166 (93%)	101 (28%)	264 (72%)	176 (85%)	30 (15%)	$\chi^2 = 285, p < .0001$	0.58
$p \geq .56$	16 (95%)	162 (91%)	120 (33%)	245 (67%)	182 (88%)	24 (12%)	$\chi^2 = 273, p < .0001$	0.58
$p \geq .61$	21 (12%)	157 (88%)	135 (37%)	230 (63%)	184 (89%)	22 (11%)	$\chi^2 = 251, p < .0001$	0.56
$p \geq .66$	24 (14%)	154 (86%)	153 (42%)	212 (58%)	187 (91%)	19 (10%)	$\chi^2 = 241, p < .0001$	0.57
$p \geq .71$	31 (17%)	147 (83%)	183 (50%)	182 (50%)	190 (92%)	16 (8%)	$\chi^2 = 219, p < .0001$	0.54
$p \geq .76$	38 (21%)	149 (79%)	206 (56%)	159 (44%)	193 (94%)	13 (6%)	$\chi^2 = 207, p < .0001$	0.52
$p \geq .81$	47 (26%)	131 (74%)	238 (65%)	127 (35%)	195 (95%)	11 (5%)	$\chi^2 = 193, p < .0001$	0.51
$p \geq .86$	66 (37%)	112 (63%)	267 (73%)	98 (27%)	197 (96%)	9 (4%)	$\chi^2 = 160, p < .0001$	0.46
$p \geq .91$	82 (46%)	96 (54%)	293 (80%)	72 (20%)	200 (97%)	6 (3%)	$\chi^2 = 144, p < .0001$	0.43
$p \geq .96$	118 (66%)	60 (34%)	337 (92%)	28 (8%)	205 (100%)	1 (0%)	$\chi^2 = 113, p < .0001$	0.36

Note: p = probability.

To minimize the high level of false-positive identification, probability of invalidity was increased in .05 increments up to .96. Across all levels, group differences in false-positive identification were present (all $p < .0001$). At the highest probability level to infer invalid performance ($p \geq .96$), MCI invalid classification (28/365) was lower than the 10% criterion commonly used to characterize acceptable false-positive PVT errors in medicolegal contexts. Nevertheless, 60/178 (34%) of early AD subjects were classified as invalid using this highly conservative threshold.

AVLT Recognition

Group results. The average delayed AVLT recognition raw score was 7.1/15 ($SD = 4.0$) for early AD, 9.8/15 ($SD = 3.6$) for MCI, and 12.9/15 ($SD = 2.7$) for controls. The values significantly differed across groups using one-way ANOVA ($p < .0001$, partial $\eta^2 = 0.26$).

Individual classification. Individual classification was performed across multiple AVLT recognition cutpoints ranging from ≤2/15 to ≤9/15, and across all thresholds, significant group differences in false-positive error rates were observed (Table 4). However, because the literature has suggested AVLT recognition cutoffs of ≤5/15 (Binder et al., 1993) and ≤9/15 for clinical use (Whitney & Davis, 2015), classification rates for these values are considered in greater detail. Across classification thresholds, the frequency of false-positive error rates differed by group membership.

Using an AVLT recognition ≤9/15 criterion, there were 18/206 (9%) false-positive controls, although 155/365 (42%) of MCI subjects and 125/178 (70%) of early AD subjects had performance levels that were considered to be invalid. Using an AVLT recognition ≤5/15 criterion, false positives for MCI were 46/365 (13%) which increased to 66/178 (37%) for early AD subjects.

We performed pairwise group follow-up comparisons using the AVLT recognition ≤9/15 and ≤5/15 thresholds. At the AVLT recognition ≤9/15 cutpoint, significant differences between controls and MCI ($\chi^2 = 70.9, p < .0001, \eta = 0.35$) as well as MCI versus early AD ($\chi^2 = 36.9, p < .0001, \eta = 0.26$) were present. When using the AVLT recognition ≤5 criterion, significant differences between controls and MCI ($\chi^2 = 8.9, p = .002, \eta = 0.12$) and MCI versus early AD ($\chi^2 = 36.9, p < .0001, \eta = 0.26$) were present.

Multiple PVT Failures

Because of the high false-positive rates observed across individual PVTs, we evaluated classification rates for subjects failing both RDS and AVLT recognition using two different thresholds for each measure. Given the extremely high false-positive rate associated with logistic regression prediction, this measure was not further investigated with pairwise combination. We used RDS cutpoints of RDS ≤ 7 and RDS ≤ 6 combined with AVLT recognition score ≤9/15 and AVLT recognition score ≤5/15.

Combining RDS ≤ 7 scores and AVLT recognition ≤9/15 performance resulted in no false positives in controls, 24/365 (7%) in MCI subjects, and 40/178 (22%) in early AD patients. Combining RDS ≤ 7 failure with a more conservative AVLT recognition ≤ 5/15 threshold decreased the false-positive rate in MCI to 9/365 (2%) and 19/178 in early AD (11%).

Combining RDS ≤ 6 failure and AVLT recognition ≤9/15 criterion resulted in no false positives in controls, 4/365 (1%) in MCI, and 9/178 (6%) in early AD. Combining RDS ≤ 6 failure with ≤5/15 AVLT recognition threshold reduced the false-positive rate in MCI to 2/365 (<1%) and 9/178 in early AD (4%).

Table 4. Classification of performance based upon AVLT recognition across cutpoints ranging from ≤9 to ≤2

	Early Alzheimer disease ($n = 178$)		Mild cognitive impairment ($n = 365$)		Controls ($n = 206$)		Significance	Effect size
AVLT recognition cutpoint	Valid	Invalid	Valid	Invalid	Valid	Invalid	χ^2	η
≤9	53 (30%)	125 (70%)	210 (58%)	155 (42%)	188 (91%)	18 (9%)	$\chi^2 = 153, p < .0001$	0.45
≤8	72 (40%)	106 (60%)	240 (66%)	125 (34%)	192 (93%)	14 (7%)	$\chi^2 = 121, p < .0001$	0.40
≤7	83 (47%)	95 (53%)	272 (74%)	93 (26%)	195 (95%)	11 (5%)	$\chi^2 = 113, p < .0001$	0.39
≤6	101 (57%)	77 (43%)	295 (81%)	70 (19%)	195 (95%)	11 (5%)	$\chi^2 = 84, p < .0001$	0.33
≤5	112 (63%)	66 (37%)	319 (87%)	46 (13%)	196 (95%)	10 (5%)	$\chi^2 = 79, p < .0001$	0.31
≤4	129 (72%)	49 (28%)	336 (92%)	29 (8%)	200 (97%)	6 (3%)	$\chi^2 = 66\, p < .0001$	0.27
≤3	142 (80%)	36 (20%)	345 (94%)	20 (6%)	203 (98%)	3 (2%)	$\chi^2 = 52, p < .0001$	0.24
≤2	152 (85%)	26 (15%)	350 (96%)	15 (4%)	204 (99%)	2 (1%)	$\chi^2 = 36, p < .0001$	0.21

Note: AVLT = Auditory Verbal Learning Test.

Predictors of PVT Failure

We performed separate multiple regression analyses to explore predictors of RDS and AVLT recognition in early AD, MCI, and controls. Similar analyses were not performed for logistic regression, given its high false-positive error rates across clinical groups. We first explored the influence of demographic predictors of age, education, and sex in each group. For early AD, age was a predictor of RDS ($p = .002$), although the total R^2 across predictors was modest ($R^2 = 0.04$). For MCI, there were no significant demographic predictors of RDS. For controls, the single significant predictor of RDS was education ($p = .02$), although overall multivariate prediction was low ($R^2 = 0.06$).

Cognitive measures used to predict RDS and AVLT recognition in the three subject groups included MMSE, AVLT LOT age-scaled score, AVLT delayed free recall age scaled score, Boston Naming age-scaled score, Animal Fluency age-adjusted z-score, Trail Making Part B age-scaled score, and Coding age-scaled score.

Reliable Digit Span. When predicting RDS in early AD, Trail Making Part B was significant ($p < .04$) with an overall multivariable $R^2 = 0.07$. When predicting RDS in MCI, Trail Making Part B was again statistically significant ($p < .0001$) as well as MMSE ($p < .006$) which was associated with an overall multivariable $R^2 = 0.12$. When predicting RDS in controls, MMSE was the single predictor ($p < .02$) with a corresponding multivariable $R^2 = 0.05$.

AVLT recognition. When predicting AVLT recognition in the early AD group, AVLT delayed free recall ($p = .006$) and AVLT LOT ($p < .03$) were both predictors and an overall $R^2 = 0.12$ was obtained. Predictors of AVLT recognition in MCI included AVLT delayed free recall ($p < .0001$), AVLT LOT ($p = .0085$), and MMSE ($p < .04$) associated with an overall $R^2 = 0.30$. When predicting AVLT recognition in controls, AVLT delayed free recall ($p < .0001$) and MMSE ($p = .009$) were predictors with overall $R^2 = 0.24$.

Classification tables. Because of the association of performance of various cognitive measures to PVT measures of RDS and AVLT recognition, classification tables reporting false-positive rates at various thresholds at different test performance levels are included.

Discussion

While PVTs have been widely investigated in medicolegal contexts and in simulator studies, less research has been performed on PVT specificity in patient series with independently established significant neurological disease. In the present project, we examine three embedded PVT measures in a large sample of research volunteers who were independently diagnosed as having either early AD ($n = 187$) or amnestic MCI ($n = 365$), or participated as cognitively intact controls ($n = 206$) based upon normal cognitive test results during screening.

All PVTs were associated with an unacceptably high level of false-positive classification in both MCI and early AD groups, and regardless of how classification thresholds were adjusted, false-positive frequency differed across groups. The error rate remained high across early AD, and it was not until PVT combined failure on both RDS and AVLT recognition that acceptable false-positive classification rates for early AD were observed.

Importantly, our data demonstrate how elevated false-positive rates can be reduced through combinations of PVTs, and by considering performance on clinical measures such as the MMSE, Trail Making Part B, and delayed AVLT free recall. Linking RDS and AVLT recognition performance led to lower false-positive rates than by using either test alone. Table 5 shows false-positive

Table 5. RDS classification across levels of MMSE scores (all groups combined) at ≤7 and ≤6 cutpoints

RDS cutpoint	Validity	MMSE = 30 (n = 117)	MMSE = 29 (n = 140)	MMSE = 28 (n = 93)	MMSE = 27 (n = 74)	MMSE = 26 (n = 95)	MMSE = 25 (n = 78)	MMSE = 24 (n = 57)	MMSE = 23 (n = 30)	MMSE = 22 (n = 21)	MMSE = 21 (n = 27)	MMSE ≤ 20 (n = 17)
≤7	Valid	109 (93%)	130 (93%)	81 (87%)	61 (82%)	75 (79%)	56 (72%)	48 (84%)	19 (63%)	10 (48%)	19 (70%)	12 (71%)
	Invalid	8 (7%)	10 (7%)	12 (13%)	13 (18%)	20 (21%)	22 (28%)	9 (16%)	11 (37%)	11 (52%)	8 (30%)	5 (29%)
≤6	Valid	115 (98%)	134 (96%)	92 (99%)	72 (98%)	86 (90%)	70 (90%)	55 (96%)	27 (90%)	15 (71%)	23 (85%)	14 (82%)
	Invalid	2 (2%)	6 (4%)	1 (1%)	2 (3%)	9 (10%)	8 (10%)	2 (4%)	3 (10%)	6 (28%)	4 (15%)	3 (18%)

Note: RDS = Reliable Digit Span; MMSE = Mini Mental Status Examination.

rates of 10% or less are obtained for RDS ≤ 6 for subjects who obtain an MMSE of 23/30 or higher. For AVLT recognition ≤5/15, false-positive rates are 15% or less for MMSE scores of at least 27/30, compared with 37% for early AD and 13% for MCI in Table 3. The false-positive rate for RDS ≤ 6 is 8% or less for subjects who produce an age-scaled score of at least 8 on Trail Making Part B. AVLT recognition ≤5/15 has a false-positive rate of 10% for subjects who score as low as an age-scaled score of 5 on AVLT delayed free recall.

RDS threshold of ≤6 misclassified 14% of dementia patients with MMSE scores of 21–30/30, increasing to 40% of patients with MMSE scores of 15–20/30, and 78% of patients with MMSE scores <15/30. A similar pattern was observed by Kiewel and colleagues (2012) who reported a false-positive error rate of 11% for the 78 subjects classified as mild AD (mean MMSE = 23/30) using an RDS criterion of ≤6. In the present study, the false-positive error rate with the ≤6 threshold yielded comparable results (13%) indicating that even with patients with mild dementia as reflected by MMSE scores that are 20/30 and above, false-positive rates for RDS likely exceed 10%.

This report extends the Kiewel and colleagues (2012) RDS findings in several important ways. First, we examine false-positive rates of RDS classification using different RDS cutpoints in a large cohort of subjects that are well characterized neurologically including a group of amnestic MCI subjects considered to have prodromal AD. This group is especially challenging because a fundamental assumption of PVT assessments is that the specificity of the PVT technique is relatively unaffected by legitimate cognitive impairment such that when poor PVT scores are obtained, insufficient task engagement can be inferred. However, poor PVT performance in patients with more subtle neurological disease might occur and thereby affect the conclusions regarding the validity of the test results. This in turn could have deleterious effects on an individual's ability to obtain needed medical and social services. Second, the current study has a robust sample size ranging from 178 to 365 subjects across groups, and includes subjects volunteering for research participation rather than clinically referred patients. Finally, we also expand the sample to capture the spectrum of cognitive aging by including persons not only with amnestic MCI, but those who are cognitively intact.

RDS is not as sensitive as other PVTs when contrasting simulators to a group of TBI patients with mean Glasgow Coma Scale scores = 9.4/15 (Bashem et al., 2014). Compared with both AVLT logistic regression and AVLT recognition classification, however, far fewer RDS false-positive invalid characterizations are made in early AD and MCI, particularly when using the classification threshold of RDS ≤ 6. This is not a surprising finding since auditory attention span is relatively unaffected by AD until the more advanced states of the disease. In contrast, impaired learning and memory are frequently seen in the early stages. Thus, high false-positive rates with memory-based approaches will likely be high in conditions in which memory impairment is a core feature.

The high false-positive error rate for logistic regression likely results from the contribution of the AVLT learning trials, given the substantially lower rate of misclassification using AVLT recognition alone. Both AD and amnestic MCI are characterized by primary impairments in verbal episodic memory, which would be much more evident for free-recall versus recognition testing. This calls for caution in this logistic regression approach in disorders involving verbal episodic memory such as AD, amnestic MCI, and dominant temporal lobe epilepsy. Using a different AVLT index that incorporates atypical patterns of recognition such as words freely recalled but not correctly recognized (Barrash, Suhr, & Manzel, 2004), a specificity of 0.94 was observed in a sample of 56 temporal lobe epilepsy patients undergoing pre-surgical evaluation (Silverberg & Barrash, 2005).

An important caveat on the generalization/validity of the logistic regression classification relates to methods and criteria for group membership used to derive the prediction equation. The sample contrasted TBI patients (18% moderate/severe) who passed all PVTs ($n = 68$) with those failing two or more PVTs ($n = 62$) identified from a series of 167 patients being evaluated for civil litigation or disability claims (Davis et al., 2012). Since all subjects had external incentive, 22% of the subject pool ($n = 37$) was excluded as being indeterminate due to failure on only a single PVT to ensure that all patients used to derived the prediction equation had unambiguous motivational status. Unfortunately, exclusion of subjects failing a single PVT introduces spectrum bias by discarding a portion of the relevant clinical sample, and fails to account for subjects not analyzed as emphasized by current reporting standards such as STROBE (Loring & Bowden, 2014). Thus, when derived PVT criteria are prospectively applied to new clinical samples, some patients will have similar performance levels as the excluded indeterminate group but in whom classification accuracy is unknown. Classifying all subjects not meeting the malingering criterion (i.e., failing at least two PVTs) as non-malingering may inappropriately characterize cases performing invalidly as valid, just as classifying the performance of all subjects failing a single PVT as malingering may inappropriately characterize valid cases as invalid, although both approaches characterize all subjects in the sample. An alternative approach is to apply the classification algorithm derived from the definite pass/definite fail group to the intermediate group that was excluded when deriving the classification formula. Doing so provides relevant information defining the boundaries of valid and invalid classification for indeterminate patients, which facilitates accurate clinical interpretation in future cases.

From our perspective, poor PVT specificity in many AD patients is not problematic since disease effects are clearly evident based upon history and activities of daily living. Thus, neuropsychological findings, when they are obtained, are primarily descriptive rather than diagnostic and are not performed in the context of external incentives, although even in these circumstances, some

clinically referred patients may be insufficiently motivated or engaged with neuropsychological findings that may underestimate true ability levels. The issue of genuine cognitive contributions to failure on PVTs, however, is addressed in common classification criteria of malingered neurocognitive impairment in which performance cannot be accounted for from neurological factors (i.e., Slick et al. Criterion D). Indeed, in describing the need for better understanding of PVT performance in dementia, motivations for feigning symptoms include competency to stand trial in criminal proceedings, in personal injury cases involving toxic exposure, or for poor medical outcomes/medical malpractice (Dean et al., 2009). A need for PVT testing in routine clinical assessment of dementia is not well articulated, and the need for accurate PVTs likely diminishes as dementia severity increases. Less reliance on PVT indicators as a function of increasing levels of dementia is analogous to other approaches in which PVT results are either discounted or completely discarded (i.e., Genuine Memory Impairment Profile; Howe & Loring, 2009). In the current context, the presence of dementia is established based upon all available clinical information and is diagnosed independently rather than relying on a component of the PVT itself to consider PVT results suspect.

These data demonstrate that specificity statistics cannot necessarily be generalized across various diseases or conditions, but rather should be empirically established. Poor specificity, as demonstrated by our MCI patients, is a serious issue in cases of milder dementia in which formal neuropsychological reports may form a primary basis for establishing disability benefits (i.e., there are external incentives to underperform). In our cognitive neurology specialty clinic at Emory University, we have evaluated multiple patients who had undergone neuropsychological testing by community psychologists who formulaically infer malingering based upon PVT scores below cutoffs from the medicolegal TBI literature. In many cases, accurate diagnosis was substantially delayed, with some patients inappropriately denied disability benefits, and in others, treatment unnecessarily postponed. Because PVTs can be influenced by cognitive impairment, knowledge of their empirically established base rate failure is necessary when used in neurologic populations. Incorrect assertion of malingering has very significant consequences for patients, both emotionally and financially, which are very difficult to reverse.

There are multiple strengths to this report. This represents the largest sample to date examining PVT in early AD, and this diagnosis was established independently from any of the primary neuropsychological measures reported here. Since these data were prospectively collected for research, it avoids the spectrum disease bias associated with clinical referral. Further, the magnitude of dementia is mild, with all early AD subjects having MMSE scores of at least 20 to be included in this study cohort. These factors also apply to our MCI subjects, a group of patients for whom PVT performance has not yet been appropriately characterized.

A disadvantage of this approach however is that research volunteers tend to be better educated than the general population (Martinson et al., 2010) and which is reflected in the estimated Verbal IQ of the sample which ranged from 120 in controls to 114 in the early AD group. Moreover, our sample was 93% Caucasian, and 61% had a college degree or higher. Because this sample reflects higher cognitive reserve, it is likely that RDS classification rates would even poorer for individuals with lower education. However, a benefit of this sample is that research volunteers are more likely to be highly motivated with little or no incentive for anything other than good faith performance during testing since their incentive for research participation is to further knowledge, which in the present context, involves MCI and early AD. Although this sample was not administered stand-alone PVT measures, only 5/749 subjects (0.6%) had lower AVLT recognition memory scores compared with delayed free recall scores, and none of these were AD subjects. Unlike clinical evaluation in which cognitive testing can be conceptualized as a bottom-up process based upon physician or spouse concerns, research subjects are self-selected and volunteer their participation and then only after complete and full informed consent reflecting a top-down approach. Unlike college students in simulator studies who may have limited commitment to participating in a single research session, higher levels of motivation can also be inferred based upon their clinical research commitment, which for ADNI is 3 years duration, a willingness to undergo repeated PET and MRI scanning, with many subjects also undergoing repeat lumbar puncture.

A limitation of this study is the use of a different AVLT recognition technique than has been used in prior reports. In the present study, the recognition format described by the MOANS cohort was employed in which the 15 target words and 15 foils are presented on a single sheet of paper and the subject circles the recognized words (Ivnik et al., 1992a, 1992b). The Davis and colleagues (2012) recognition format included a list of 50 words read orally to the subject with the subject sequentially indicating whether or not each word was on the original list. Thus, our recognition approach differed in both modality of presentation and in number of distractor items. However, because the number of words correctly recognized is the dependent measure rather than a corrected recognition score that included a correction for incorrect recognitions of non-target foils, the effect of this difference is expected to be small. If present, we would expect that fewer words would bias the results toward the null since with fewer words as distractors, each target word has greater salience.

These data address the issue of false-positive classification alone. We did not have a separate litigating sample failing multiple PVTs without evidence of MCI or early AD, nor did we have a sample of normal subjects asked to feign impairment. Without these comparison groups, we could not address the effects on sensitivity to invalid performance caused by improving specificity in MCI and early AD. There is always a tradeoff between improving specificity and lowering sensitivity; as one improves the other declines, and vice versa.

Table 6. RDS classification across levels of Trail Making Part B age-scaled scores (all groups combined) at ≤7 and ≤6 cutpoints

RDS cutpoint	Validity	SS ≥ 18 ($n = 2$)	SS = 17 ($n = 6$)	SS = 16 ($n = 16$)	SS = 15 ($n = 27$)	SS = 14 ($n = 65$)	SS = 13 ($n = 54$)	SS = 12 ($n = 61$)	SS = 11 ($n = 51$)	SS = 10 ($n = 124$)
≤7	Valid	2 (100%)	5 (83%)	15 (94%)	27 (100%)	63 (97%)	46 (85%)	57 (93%)	46 (90%)	106 (86%)
	Invalid	0 (0%)	1 (17%)	1 (6%)	0 (0%)	2 (3%)	8 (15%)	4 (7%)	5 (10%)	18 (14%)
≤6	Valid	32 (100%)	5 (83%)	16 (100%)	27 (100%)	64 (98%)	52 (96%)	60 (98%)	50 (98%)	120 (97%)
	Invalid	0 (0%)	1 (17%)	0 (0%)	0 (0%)	1 (2%)	2 (4%)	1 (2%)	1 (2%)	4 (3%)
RDS cutpoint	Validity	SS = 9 ($n = 75$)	SS = 8 ($n = 38$)	SS = 7 ($n = 75$)	SS = 6 ($n = 14$)	SS = 5 ($n = 5$)	SS = 4 ($n = 8$)	SS = 3 ($n = 20$)	SS = 2 ($n = 108$)	
≤7	Valid	64 (85%)	32 (84%)	58 (77%)	9 (64%)	1 (20%)	8 (100%)	12 (60%)	69 (64%)	
	Invalid	11 (15%)	6 (16%)	17 (23%)	5 (36%)	4 (80%)	0 (0%)	8 (40%)	39 (36%)	
≤6	Valid	71 (95%)	35 (92%)	67 (89%)	13 (93%)	1 (20%)	8 (100%)	18 (90%)	91 (84%)	
	Invalid	4 (5%)	3 (8%)	8 (11%)	1 (7%)	4 (80%)	0 (0%)	2 (10%)	17 (16%)	

Note: RDS = Reliable Digit Span; SS=scaled score.

Table 7. RDS classification across levels of AVLT delayed free recall age scaled scores (all groups combined) at ≤7 and ≤6 cutpoints

RDS cutpoint	Validity	SS = 18 (n = 11)	SS = 17 (n = 11)	SS = 16 (n = 7)	SS = 15 (n = 8)	SS = 14 (n = 30)	SS = 13 (n = 26)	SS = 12 (n = 26)	SS = 11 (n = 44)	SS = 10 (n = 42)
≤7	Valid	11 (100%)	10 (91%)	4 (57%)	8 (100%)	26 (87%)	22 (85%)	20 (77%)	43 (98%)	39 (93%)
	Invalid	0 (0%)	1 (9%)	3 (43%)	0 (0%)	4 (13%)	4 (15%)	6 (23%)	1 (2%)	3 (7%)
≤6	Valid	11 (100%)	10 (91%)	6 (86%	8 (100%)	29 (97%)	26 (100%)	23 (88%)	44 (100%)	41 (98%)
	Invalid	0 (0%)	1 (9%)	1 (14%)	0 (0%)	1 (3%)	0 (0%)	3 (12%)	0 (0%)	1 (2%)
RDS cutpoint	Validity	SS = 9 (n = 58)	SS = 8 (n = 38)	SS = 7 (n = 78)	SS = 6 (n = 127)	SS = 5 (n = 81)	SS = 4 (n = 73)	SS = 3 n = 63)	SS = 2 (n = 26)	
≤7	Valid	48 (83%)	32 (84%)	69 (88%)	108 (85%)	59 (73%)	59 (81%)	44 (70%)	18 (69%)	
	Invalid	10 (17%)	6 (16%)	9 (12%)	19 (15%)	22 (27%)	14 (19%)	19 (30%)	8 (31%)	
≤6	Valid	54 (93%)	34 (90%)	75 (96%)	122 (96%)	73 (90%)	69 (94%)	56 (89%)	22 (85%)	
	Invalid	4 (7%)	4 (10%)	3 (4%)	5 (4%)	8 (10%)	4 (6%)	7 (11%)	4 (15%)	

Note: AVLT = Auditory Verbal Learning Test; RDS = Reliable Digit Span, SS = scaled score.

Table 8. AVLT recognition classification across levels of MMSE scores (all groups combined) at ≤9 and ≤5 cutpoints

AVLT recognition cutpoint	Validity	MMSE = 30 (n = 117)	MMSE = 29 (n = 140)	MMSE = 28 (n = 93)	MMSE = 27 (n = 74)	MMSE = 26 (n = 95)	MMSE = 25 (n = 78)	MMSE = 24 (n = 57)	MMSE = 23 (n = 30)	MMSE = 22 (n = 21)	MMSE = 21 (n = 27)	MMSE ≤ 20 (n = 17)
≤9	Valid	97 (83%)	115 (82%)	64 (69%)	44 (60%)	51 (54%	34 (43%)	27 (47%)	6 (20%)	3 (14%)	7 (26%)	14 (82%)
	Invalid	20 (17%)	25 (18%)	29 (31%)	30 (40%)	44 (46%)	44 (56%)	30 (53%)	24 (80%)	18 (86%)	20 (74%)	3 (18%)
≤5	Valid	108 (92%)	136 (97%)	86 (92%)	63 (85%)	73 (77%)	63 (81%)	45 (79%)	19 (63%	7 (33%)	18 (67%)	9 (53%)
	Invalid	9 (8%)	4 (3%)	7 (8%)	11 (15%)	22 (23%)	15 (19%)	12 (21%)	11 (37%)	14 (67%)	9 (33%)	8 (47%)

Note: AVLT = Auditory Verbal Learning Test; MMSE = Mini Mental Status Examination.

Table 9. AVLT recognition classification across levels of Trail Making Part B age-scaled scores (all groups combined) at ≤9 and ≤5 cutpoints

AVLT recognition cutpoint	Validity	SS ≥ 18 ($n = 2$)	SS = 17 ($n = 6$)	SS = 16 ($n = 16$)	SS = 15 ($n = 27$)	SS = 14 ($n = 65$)	SS = 13 ($n = 54$)	SS = 12 ($n = 61$)	SS = 11 ($n = 51$)	SS = 10 ($n = 124$)
≤9	Valid	2 (100%)	6 (100%)	11 (69%)	20 (74%)	55 (85%)	43 (80%)	45 (74%)	37 (72%)	78 (63%)
	Invalid	0 (0%)	0 (0%)	5 (31%)	7 (26%)	10 (15%)	11 (20%)	16 (26%)	14 (28%)	46 (37%)
≤5	Valid	2 (100%)	6 (100%)	15 (94%)	25 (93%)	63 (97%)	49 (91%)	53 (87%)	44 (86%)	109 (88%)
	Invalid	0 (0%)	0 (0%)	1 (6%)	2 (7%)	2 (3%)	5 (9%)	8 (13%)	7 (14%)	15 (12%)
AVLT recognition cutpoint	Validity	SS = 9 ($n = 75$)	SS = 8 ($n = 38$)	SS = 7 ($n = 75$)	SS = 6 ($n = 14$)	SS = 5 ($n = 5$)	SS = 4 ($n = 8$)	SS = 3 ($n = 20$)	SS = 2 ($n = 108$)	
≤9	Valid	41 (55%)	22 (58%)	36 (48%)	5 (36%)	2 (40%)	3 (38%)	6 (30%)	39 (36%)	
	Invalid	34 (45%)	16 (42%)	39 (52%)	9 (84%)	3 (60%)	5 (62%)	14 (70%)	69 (64%)	
≤5	Valid	61 (81%)	33 (87%)	58 (77%)	7 (50%)	4 (80%)	6 (75%)	13 (65%)	79 (73%)	
	Invalid	14 (19%	5 (13%)	17 (23%)	7 (50%)	2 (20%)	2 (25%)	7 (35%)	29 (27%)	

Note: AVLT = Auditory Verbal Learning Test; SS = scaled score.

Table 10. AVLT recognition classification across levels of AVLT delayed free recall age-scaled scores (all groups combined) at ≤9 and ≤5 cutpoints

AVLT recognition cutpoint	Validity	SS = 18 (n = 11)	SS = 17 (n = 11)	SS = 16 (n = 7)	SS = 15 (n = 8)	SS = 14 (n = 30)	SS = 13 (n = 26)	SS = 12 (n = 26)	SS = 11 (n = 44)	SS = 10 (n = 42)
≤9	Valid	11 (100%)	11 (100%)	7 (100%)	8 (100%)	29 (97%)	25 (96%)	24 (92%)	41 (93%)	37 (88%)
	Invalid	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (3%)	1 (4%)	2 (7%)	3 (7%)	5 (12%)
≤5	Valid	11 (100%)	11 (100%)	7 (100%)	8 (100%)	29 (97%)	25 (96%)	26 (100%)	44 (100%)	40 (95%)
	Invalid	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (3%)	1 (4%)	0 (0%)	0 (0%)	2 (5%)
AVLT recognition cutpoint	Validity	SS = 9 (n = 58)	SS = 8 (n = 38)	SS = 7 (n = 78)	SS = 6 (n = 127)	SS = 5 (n = 81)	SS = 4 (n = 73)	SS = 3 (n = 63)	SS = 2 (n = 26)	
≤9	Valid	49 (84%)	28 (74%)	50 (64%)	58 (46%)	30 (37%)	16 (22%)	22 (35%)	5 (19%)	
	Invalid	9 (16%)	10 (26%)	28 (36%)	69 (54%)	51 (63%)	57 (78%)	41 (65%)	21 (81%)	
≤5	Valid	55 (95%)	36 (95%)	67 (86%)	98 (77%)	61 (75%)	50 (68%)	44 (70%)	15 (58%)	
	Invalid	3 (5%)	2 (5%)	11 (14%)	29 (32%)	20 (25%)	23 (32%)	19 (20%)	11 (42%)	

Note: AVLT = Auditory Verbal Learning Test; SS = scaled score.

These results demonstrate the importance of cross-validating PVTs, not only on independent samples that are similar to the initial validation study, but also on samples of subjects with significant neurologic, psychiatric, or developmental disorders who are not in settings with external incentives to underperform. This is necessary not only to identify risk factors for false-positive identification, but also to establish PVT modifications/adaptations needed to reduce the likelihood of misinterpreting performance on PVTs as invalid when, in fact, the performance accurately represents an examinee's true ability level. As we have demonstrated, combinations of PVTs in early AD and MCI can reduce the per-test false-positive rate. The false-positive rate can also be lowered by considering ability level as reflected by global cognitive status, processing speed, and delayed free recall, and considering whether a patient has sufficient cognitive resources to pass the PVT (see Tables 5–10). In our opinion, research minimizing false-positive errors represents the next wave of research on PVTs, as evidenced by recent papers on false-positive error rate associated with multiple PVT use (Bilder, Sugar, & Hellemann, 2014; Davis & Millis, 2014; Larrabee, 2014).

Funding

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

This work was also supported by the Emory Alzheimer's Disease Research Center (NIH-NIA 5 P50 AG025688).

Conflict of interest

None declared.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Aisen, P. S., Petersen, R. C., Donohue, M., Gamst, A., Raman, R., Thomas, R. G., et al. Alzheimer's Disease Neuroimaging Initiative. (2010). Clinical core of the Alzheimer's Disease Neuroimaging Initiative: Progress and plans. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *6 (3)*, 239–246.

Army Individual Test Battery. (1944). *Manual of directions and scorings*. Washington, DC: U.S. War Department, Adjutant General's Office.

Barrash, J., Suhr, J., & Manzel, K. (2004). Detecting poor effort and malingering with an expanded version of the Auditory Verbal Learning Test (AVLTX): Validation with clinical samples. *Journal of Clinical and Experimental Neuropsychology*, *26 (1)*, 125–140.

Bashem, J. R., Rapport, L. J., Miller, J. B., Hanks, R. A., Axelrod, B. N., & Millis, S. R. (2014). Comparisons of five performance validity indices in bona fide and simulated traumatic brain injury. *The Clinical Neuropsychologist*, *28 (5)*, 851–875.

Bilder, R. M., Sugar, C. A., & Hellemann, G. S. (2014). Cumulative false positive rates given multiple performance validity tests: Commentary on Davis and Millis (2014) and Larrabee (2014). *The Clinical Neuropsychologist*, *28 (8)*, 1212–1223.

Binder, L. M., Villanueva, M. R., Howieson, D., & Moore, R. T. (1993). The Rey AVLT recognition memory task measures motivational impairment after mild head trauma. *Archives of Clinical Neuropsychology*, *8*, 137–147.

Bondi, M. W., & Smith, G. E. (2014). Mild cognitive impairment: A concept and diagnostic entity in need of input from neuropsychology. *Journal of the International Neuropsychological Society*, *20 (2)*, 129–134.

Boone, K. B. (2013). *Clinical practice of forensic neuropsychology*. New York, NY: Guilford.

Boone, K. B., Lu, P., & Wen, J. (2005). Comparison of various RAVLT scores in the detection of noncredible memory performance. *Archives of Clinical Neuropsychology*, *20 (3)*, 301–319.

Bush, S. S., Ruff, R. M., Tröster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., et al. (2005). Symptom validity assessment: Practice issues and medical necessity NAN policy & planning committee. *Archives of Clinical Neuropsychology*, *20 (4)*, 419–426.

Davis, J. J., & Millis, S. R. (2014). Examination of performance validity test failure in relation to number of tests administered. *The Clinical Neuropsychologist*, *28 (2)*, 199–214.

Davis, J. J., Millis, S. R., & Axelrod, B. N. (2012). Derivation of an embedded Rey Auditory Verbal Learning Test performance validity indicator. *The Clinical Neuropsychologist*, *26 (8)*, 1397–1408.

Dean, A. C., Victor, T. L., Boone, K. B., Philpott, L. M., & Hess, R. A. (2009). Dementia and effort test performance. *The Clinical Neuropsychologist*, *23 (1)*, 133–152.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state." *Journal of Psychiatric Research*, *12*, 189–198.

Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment*, *6 (3)*, 218–224.

Grober, E., & Sliwinski, M. (1991). Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *Journal of Clinical & Experimental Neuropsychology*, *13 (6)*, 933–949. doi: 10.1080/01688639108405109

Hachinski, V. C., Iliff, L. D., Zilhka, E., Du Boulay, G. H., McAllister, V. L., Marshall, J., et al. (1975). Cerebral blood flow in dementia. *Archives of Neurology*, *32 (9)*, 632–637.

Heinly, M. T., Greve, K. W., Bianchini, K. J., Love, J. M., & Brennan, A. (2005). WAIS digit span-based indicators of malingered neurocognitive dysfunction: Classification accuracy in traumatic brain injury. *Assessment*, *12 (4)*, 429–444.

Howe, L. L., & Loring, D. W. (2009). Classification accuracy and predictive ability of the medical symptom validity test's dementia profile and general memory impairment profile. *The Clinical Neuropsychologist*, *23 (2)*, 329–342.

Hughes, C. P., Berg, L., Danziger, W. L., Coben, L. A., & Martin, R. L. (1982). A new clinical scale for the staging of dementia. *British Journal of Psychiatry*, *140*, 566–572.

Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kikmen, E., et al. (1992a). Mayo's older Americans normative studies: Updated AVLT Norms for ages 56–97. *The Clinical Neuropsychologist*, *6(Suppl.)*, 83–104.

Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kikmen, E., et al. (1992b). Mayo's older Americans normative studies: WAIS-R norms for ages 56–97. *The Clinical Neuropsychologist*, *6*, 1–30.

Kaplan, E. F., Goodglass, H., & Weintraub, S. (1983). *The Boston naming test*. Philadelphia: Lea & Febiger.

Kiewel, N. A., Wisdom, N. M., Bradshaw, M. R., Pastorek, N. J., & Strutt, A. M. (2012). A retrospective review of digit span-related effort indicators in probable Alzheimer's disease patients. *The Clinical Neuropsychologist*, *26 (6)*, 965–974.

Larrabee, G. J. (2003). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *The Clinical Neuropsychologist*, *17 (3)*, 410–425.

Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society*, *18 (4)*, 625–631.

Larrabee, G. J. (2014). Minimizing false positive error with multiple performance validity tests: Response to Bilder, Sugar, and Hellemann (2014). *The Clinical Neuropsychologist*, *28 (8)*, 1230–1242.

Loring, D. W., & Bowden, S. C. (2014). The STROBE statement and neuropsychology: Lighting the way toward evidence-based practice. *The Clinical Neuropsychologist*, *28 (4)*, 556–574.

Martinson, B. C., Crain, A. L., Sherwood, N. E., Hayes, M. G., Pronk, N. P., & O'Connor, P. J. (2010). Population reach and recruitment bias in a maintenance RCT in physically active older adults. *Journal of Physical Activity and Health*, *7 (1)*, 127–135.

Merten, T., Bossink, L., & Schmand, B. (2007). On the limits of effort testing: Symptom validity tests and severity of neurocognitive symptoms in nonlitigant patients. *Journal of Clinical & Experimental Neuropsychology*, *29 (3)*, 308–318.

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). New York: Oxford University Press.

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology*, *43 (11)*, 2412–2414.

Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique. *Archives de Psychologie*, *28*, 286–340.

Rogers, R. (1997). Researching dissimulation. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 398–426). New York, NY: Guilford Press.

Schmidt, M. (1996). *Rey auditory and verbal learning test: A handbook*. Los Angeles: Western Psychological Services.

Silverberg, N., & Barrash, J. (2005). Further validation of the expanded auditory verbal learning test for detecting poor effort and response bias: Data from temporal lobectomy candidates. *Journal of Clinical & Experimental Neuropsychology*, *27 (7)*, 907–914.

Slick, D. J., Sherman, E. M., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist*, *13 (4)*, 545–561.

Steinberg, B. A., Bieliauskas, L. A., Smith, G. E., & Ivnik, R. J. (2005). Mayo's Older Americans Normative Studies: Age- and IQ-adjusted norms for the Trail-Making Test, the Stroop Test, and MAE Controlled Oral Word Association test. *The Clinical Neuropsychologist*, *19 (3–4)*, 329–377.

Steinberg, B. A., Bieliauskas, L. A., Smith, G. E., Langellotti, C., & Ivnik, R. J. (2005). Mayo's Older Americans Normative Studies: Age- and IQ-Adjusted Norms for the Boston Naming Test, the MAE Token Test, and the Judgment of Line Orientation test. *The Clinical Neuropsychologist*, *19 (3–4)*, 280–328.

Straus, S. E., Richardson, W. S., Glasziou, P., & Haynes, R. B. (2010). *Evidence-based medicine: How to practice and teach EBM* (4th ed.). Edinburgh: Elsevier Churchill-Livingstone.

Tierney, M. C., Nores, A., Snow, W. G., Fisher, R. H., Zorzitto, M. L., & Reid, D. W. (1994). Use of the Rey Auditory Verbal Learning Test in differentiating normal aging from Alzheimer's and Parkinson's dementia. *Psychological Assessment*, *6 (2)*, 129–134.

Wechsler, D. (1981). *Wechsler adult intelligence scale-revised*. New York: The Psychological Corporation.
Wechsler, D. (1987). *Wechsler memory scale—revised manual*. San Antonio, TX: The Psychological Corporation.
Whitney, K. A., & Davis, J. J. (2015). The non-credible score of the Rey Auditory Verbal Learning Test: Is it better at predicting non-credible neuropsychological test performance than the RAVLT recognition score? *Archives of Clinical Neuropsychology*, *30 (2)*, 130–138.
Wolfe, P. L., Millis, S. R., Hanks, R., Fichtenberg, N., Larrabee, G. J., & Sweet, J. J. (2010). Effort indicators within the California Verbal Learning Test-II (CVLT-II). *The Clinical Neuropsychologist*, *24 (1)*, 153–168.